G. Marcou[1]
G. Delouis[1]
D. Horvath[1]
A. Varnek[1]

TRANSDUCTION LEARNING IN QSAR: AN EFFICIENT WAY TO BUILD THE MODELS ON SMALL DATA SETS

[1] Laboratoire de Chemoinformatique, UMR7140 CNRS-Université de Strasbourg, France;

g.marcou@unistra.fr

Prediction performance of conventional QSAR models built on small and diverse data sets is often very limited. This could be significantly improved using Transduction and Semi-Supervised Learning algorithms. The idea is to use at the training stage both compounds with known property values, as in conventional QSAR, and target compounds for which the property should be assessed[1]. Up to now, only few examples of the use of semi-supervised methods in computational chemistry were reported in the literature (e.g., see paper by Kondratovitch et al.[2]).

In this presentation we discuss the Transductive Ridge Regression method and its implementation in the ISIDA software package. Developed tools have been applied to 3 datasets: aqueous solubility, proteolytic ionization constants and affinity to A2AR for 1635, 924 and 767 compounds respectively, using ISIDA fragment descriptors[3] or MOE 2D descriptors[4].

The transductive effect (TE) was calculated as a difference between Balanced Accuracy parameters obtained with transductive and classical Ridge Regression algorithms. The TE was observed in more than 90% of the calculations, mostly in case of small training sets. This confirms that transduction algorithms are particularly useful in situations where the data are expensive to measure or difficult to collect.

1. Chapelle O. et al. *Semi-supervised Learning*, MIT Press: 2010.
2. Kondratovich E. et al. *Molecular Informatics*, 2013, **32**: 261-266.
3. Varnek, A. et al. *Journal of computer-aided molecular design*, 2005, **19:** 693-703.
4. *Molecular Operating Environment (MOE)*, H3A 2R7; Chemical Computing Group Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, 2017.